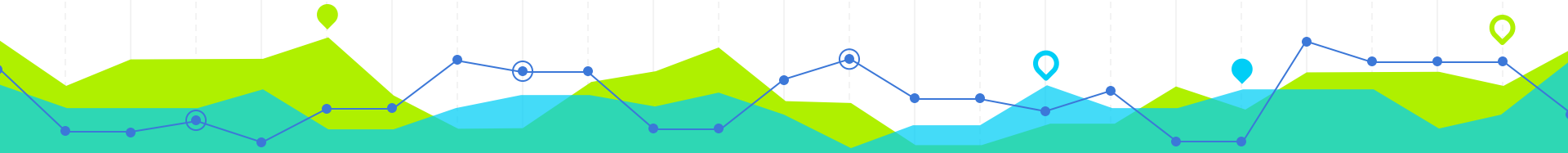
Association Rules for Fraud Detection

Mike Meissner, Joey Gaule, Adela Cho, Joseph Strickland



Background 1

MOTIVATION OF THE STUDY, OBJECTIVE, & HOW IT WAS ACCOMPLISHED

- Motivation: Businesses lose 5% of revenue (~$140,000) to fraud every year
  - Typical fraud lasts 18 months before detection & budgetary cutbacks make finding fraud difficult
- Objective: “automated or machine-driven techniques for finding fraud without expending large amounts of auditor time are valuable”
- Association Rules (AR) - data mining technique that excels in detecting transaction anomalies among millions of records
  - Provides recommendation to auditor quickly & efficiently with minimal human effort

What are Association Rules?

- Fundamental concept in data mining used to find interesting correlations, frequent patterns or associations among sets of items usually found in transaction databases
- Searches data for frequent if-then patterns and uses criterion like Support, Confidence and Lift to determine the most important relationships
- Support = how frequently an itemset appears in the data
- Confidence = measure of reliability of an inference made by a rule
- Lift = compares the strength of the association between two items to the expected strength of the association if the items were independent

$$\{\text{onions}, \text{potatoes}\} \Rightarrow \{\text{burger}\}$$

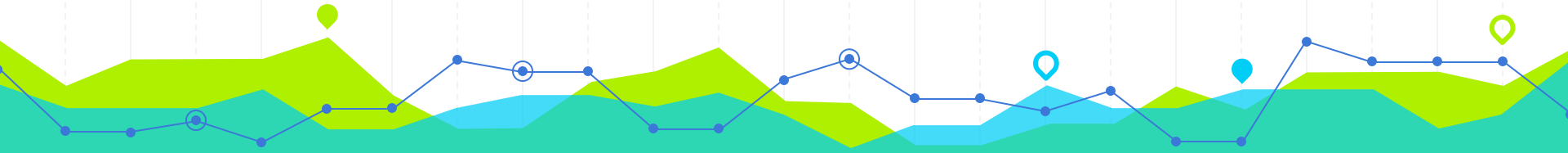
BUSINESS RELEVANCE

- Automated methods for finding fraud decreases time internal auditors have to spend on catching perpetrators
- Improve company's security measures
  - Prevent hundred thousands of revenue loss every year from fraud
  - Increase customer trust by preventing harm from loss
- Competitive advantage by building a reputation of protecting company assets and customer data

Data Overview

- The study utilized digital transaction records in CSV format from an unnamed retail store during the fourth quarter of the year
- Variables Included: Store, Department, Inventory ID, Transaction, Cash or Credit Used, Dollar Amount (Transformed into Categorical Variable based on 150$ increments), Employee ID
- Each observation represents a single transaction
- Dataset aims to generalize to all transactions for the store

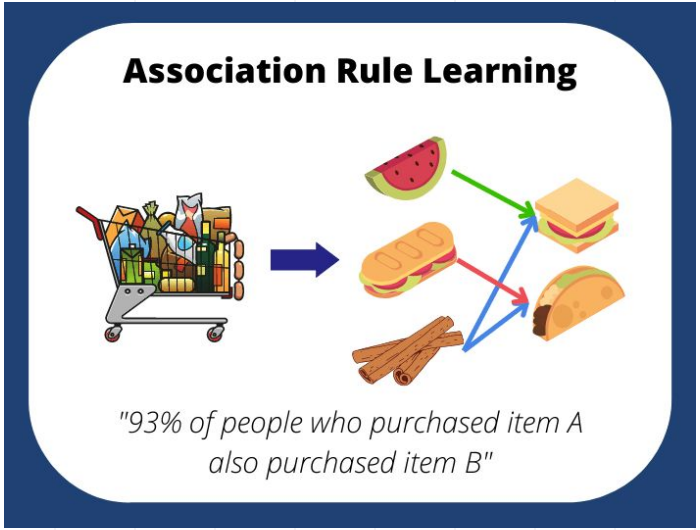
Store	*Department*	*Inventory ID*	*Transaction*	*Type*	*Size*	*Employee ID*
21	Lawn & Garden	24966	Sale	Credit	<1350	161
28	Men's Clothing	41235	Sale	Cash	<150	248
11	Toys	70730	Sale	Cash	<150	66
5	Men's Clothing	83295	Refund	Credit	<150	413
15	Lawn & Garden	55699	Sale	Credit	<450	572
22	Electronics	86682	Sale	Credit	<1050	466
16	Electronics	37680	Refund	Cash	<300	99
22	Women's Clothing	42997	Sale	Credit	<300	230
6	Men's Clothing	6751	Sale	Credit	<150	12
7	Pets	66276	Sale	Cash	<150	52



METHODOLOGY 2

STEPS OF AR (1/4)

- Identify the key variables of interest (ex. Customer refunds)
- Ask question: “Is there some event (or combination of events) that accompanies a routine customer refund that can be used to systematically flag transactions for review by internal auditors?
  - Significantly different transaction flagged
- Internal auditors submit request to IT department for flat file with relevant information



STEPS OF AR (2/4)

- Terms needed to use AR and understand output: premises, conclusion, support, confidence, & lift
  - Association rules are structure as if(premises)-then(conclusions) statements
    - Premise implies conclusion
  - Minimum support and minimum confidence are chosen by auditor as limiters
    - Expressed as percentages
    - Support shows prevalence of given relationship in data
    - Confidence is the likelihood or probability
    - Auditors want to set minimum support and minimum confidence to low level to avoid discarding potentially useful rules
  - Lift is a measure of interestingness for identifying potentially fraudulent transactions
    - High life & high confidence = needs further investigation

STEPS OF AR (3/4)

- Begin analysis using “Association Rules” in data-mining menu in software program
- Select fraud variable of interest & other related variable
- Specify minimum support and confidence levels
- Start program and let it calculate all AR for defined variables that meet min support and confidence specifications
- Output file is a standard worksheet

$$Rule:\ X \Rightarrow Y$$

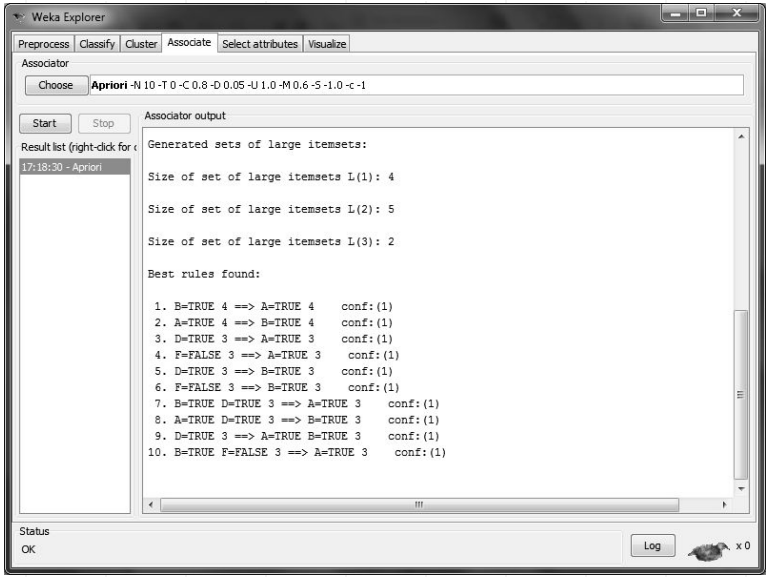
$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

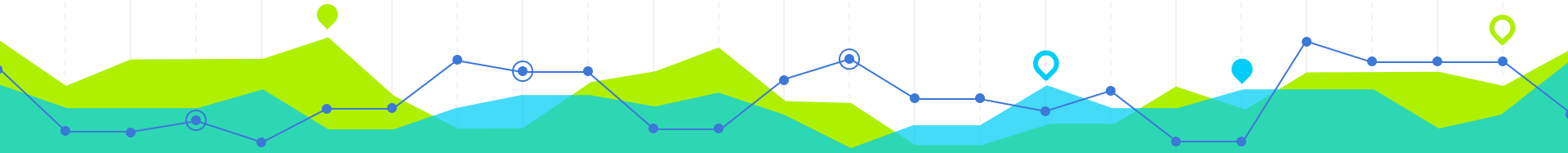
STEPS OF AR (4/4)

- Sort column labeled “Conclusion” and delete all rows except the fraud variable of interest
  - Removes 90% of AR
  - Left with various combinations of variables (premises) that imply fraud variables of interest (conclusion), with a specified degree of confidence
- Sort remaining data in table based on life in descending order
  - Puts interesting relationships to top of the table



WHY THIS METHOD?

- Discover interesting relationships between variables in large databases in a timely manner to make informed decision-making
  - Uncovers hidden patterns about the underlying structure of the data
- Allows corporations to take proactive measures to prevent fraud upon identifying patterns of fraudulent activity



APPLICATION 3

WHAT ARE THE IMPLICATIONS FOR SPECIFIC DOMAIN AREAS?(1/2)

Market Basket Analysis (MBA)

- **Description:** Uncover hidden patterns in customer purchase behavior to improve product placement, promotions, and marketing strategies.
- **Example:** Identify items frequently bought together to suggest complementary purchases at checkout or run targeted promotions for specific product combinations.

Intelligent Transportation Systems (ITS)

- **Description:** Analyze traffic data to optimize traffic flow, reduce congestion, and improve overall transportation efficiency.
- **Example:** Predict traffic patterns and suggest alternate routes to drivers, dynamically adjust traffic light timings, and identify areas for infrastructure improvements.

Web Log Analysis

- **Description:** Understand user behavior on websites to improve website design, content personalization, and targeted advertising.
- **Example:** Identify popular website sections, recommend relevant content to users based on their browsing history, and target advertising campaigns based on user interests.

WHAT ARE THE IMPLICATIONS FOR SPECIFIC DOMAIN AREAS?(2/2)

Disease Identification in Healthcare

- **Description:** Discover relationships between symptoms and diseases to aid in diagnosis and treatment planning.
- **Example:** Identify clusters of frequently co-occurring symptoms to suggest potential diagnoses, predict the risk of developing specific diseases for individuals based on their medical history.

Computer-Aided Diagnosis (CAD) Systems

- **Description:** Develop AI systems that assist doctors in diagnosing diseases by analyzing medical images and data.
- **Example:** Identify suspicious lesions in mammograms to aid in breast cancer diagnosis, analyze genetic data to predict the risk of developing specific diseases.

OTHER DOMAINS THAT MAY BENEFIT FROM THESE TOOLS? WHY?

Personalized Learning Systems:

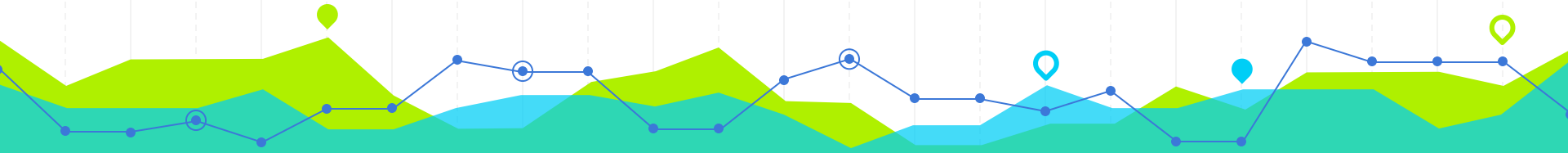
- **Description:** Analyze student learning data to personalize educational content, recommend learning resources, and predict student performance.
- **Example:** Identify learning patterns among students and tailor content difficulty and delivery methods accordingly, recommend additional resources based on individual strengths and weaknesses, predict students at risk of falling behind and provide targeted interventions.

Sustainability and Environmental Management:

- **Description:** Analyze data on energy consumption, resource use, and environmental factors to identify opportunities for sustainability improvements.
- **Example:** Discover relationships between energy consumption patterns and weather conditions to optimize energy usage based on historical data and sensor readings.

Cybersecurity Threat Detection:

- **Description:** Analyze network traffic data and system logs to identify patterns indicative of cyberattacks and security breaches.
- **Example:** Detect suspicious network activity by identifying unusual patterns in data transfer or communication protocols, predict potential security vulnerabilities based on system configurations and software versions, identify groups of compromised devices based on shared network activity patterns.



SUGGESTIONS 4

SAMPLING LIMITATIONS

- **Overview of Limitations:** The study's sampling strategy may not adequately represent the target population, leading to potential biases in the findings.

Detailed Suggestions:

- Stratified Random Sampling: Implement stratified random sampling to ensure all subgroups in the population are adequately represented, reducing sampling bias.
- Snowball Sampling for Hard-to-Reach Populations: In cases where certain population segments are difficult to access, snowball sampling can help in reaching these groups, although it's important to consider the potential for increased bias.
- Increase Sample Size: Enlarging the sample size can improve the study's power and the representativeness of the sample, especially for populations with high variability.

VALIDATE LIMITATIONS

- **Overview of Limitations:** The tools and instruments used for data collection might not have undergone sufficient validation, questioning the reliability of the data.

- **Detailed Suggestions:**
  - Cross-Validation with Existing Instruments: Where possible, validate new instruments against those already established to ensure reliability and validity.
  - Pilot Studies: Conduct pilot studies to test the instruments in a smaller, controlled setting, allowing for adjustments before the full-scale study.
  - Expert Review: Engage subject matter experts to review the instruments for content validity, ensuring they accurately capture the constructs of interest.

EXPERIMENTAL DESIGN CONSIDERATIONS

- **Overview of Limitations:** The initial experimental setup may not fully account for external variables that could influence the outcomes, potentially compromising the study's internal validity.

- **Detailed Suggestions:**
  - Control Groups: Utilize multiple control groups to cover different aspects of potential interference, ensuring that the effect is genuinely attributable to the variable of interest.
  - Random Assignment: Apply random assignment to treatment and control groups to minimize pre-existing differences between the groups.
  - Blinding: Implement single or double-blind designs wherever feasible to reduce bias from participants and researchers.

OMITTED VARIABLE BIAS

- **Overview of Limitations:** Failing to include all relevant variables in the study's design can lead to omitted variable bias, where the effect of unobserved variables skews the results.

- **Detailed Suggestions:**
  - Thorough Literature Review: Conduct an exhaustive literature review to identify all potential variables that could influence the study's outcomes.
  - Statistical Techniques: Employ statistical techniques such as regression analysis to control for the potential influence of omitted variables.
  - Sensitivity Analysis: Perform sensitivity analysis to assess how sensitive the results are to changes in the model's specifications, including omitted variables.

ALTERNATIVE METHODS

- **Overview of Limitations:** Relying solely on one methodological approach may not capture the study's full scope, missing out on nuanced insights.

- **Detailed Suggestions:**
  - **Mixed Methods Approach:** Incorporate both qualitative and quantitative methods to leverage the strengths of each. While quantitative methods provide scalability and generalizability, qualitative methods offer depth and context.
  - **Triangulation:** Use multiple data sources, investigators, and methodologies to cross-verify results, enhancing the study's credibility.
  - **Innovative Data Collection Methods:** Explore the use of novel data collection methods, such as digital trace data or ecological momentary assessment, to capture real-time data and reduce recall bias.

“

"Without data, you're just another person with an opinion." *– W. Edwards Deming*

THANKS!

Any questions?